International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.2 (2016), pp.173-184
http://dx.doi.org/10.14257/ijmue.2016.11.2.18

Support Vector Machine Prediction Model Based on Chaos Theory

Song Liangong[1], Wu Huixin[2] and Zhang Zezhong[3]

1. College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou Henan, 450011 China
2. College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou Henan, 450011 China
3. School of Water Conservancy, North China University of Water Resources and Electric Power, Zhengzhou Henan, 450011 China
E-mail

Abstract

In order to enhance prediction precision of online public opinion, it put forward a kind of online public opinion prediction model (PSR-SVR) with the combination of chaos theory and support vector regression. First of all, the original data of online public opinion were obtained throughout topic segmentation, hotspot extraction, and data aggregate. Then, time sequence of online public opinion was reconstructed throughout phase-space reconstruction. Finally, the reconstructed time sequence of online public opinion was input support vector regression for modeling and prediction, and then it was compared with other online public opinion prediction model by experiment. The result shows that compared with the contrast model, PSR-SVR improves the prediction precision and reliability of online public opinion, and the prediction results have certain practical value.

Keywords*: Online public opinion; Support vector regression; Phase space reconstruction; Chaos theory*

1. Introduction

Online public opinion is an important part of social public opinion; compared with the traditional news media, it's highly interactive. The user is not only the information receiver, but also the information source, which makes the information more timely and quickly spread on the Internet. Negative online public opinion will bring a greater threat to social and public security, so the analysis and modeling of the changes of online public opinion, and forecasting its developing trend can help the relevant departments to formulate the correct to guide public opinion, and has important practical significance to maintain social harmony and stability [1-2].

Current prediction method of online public opinion trend is mainly divided into two categories: based on the linear prediction method and based on machine learning methods; linear prediction method has autoregressive (AR), moving average (MA), difference autoregressive integrated moving average (ARIMA), *etc.* [3-5]. Such kind of methods is simple and easy to implement, especially the ARIMA is extremely flexible, and it can represent all kinds of time sequence models, combines the advantages of time sequence analysis and regression analysis, and used the most widely in online public opinion change trend prediction. However, ARIMA is a linear prediction model, but the online public opinion changes are influenced by many factors, with nonlinear characteristics. ARIMA is unable to capture the nonlinear change characteristics of online public opinion changes, so as to affect the prediction precision [6]. Machine

ISSN: 1975-0080 IJMUE

learning algorithm mainly has hidden Markova model (HMM), grey theory (GM), artificial neural network (ANN), support vector regression (SVR) and so on; this kind of methods is based on the nonlinear modeling theory, and can more accurately describe the online public opinion change trend. Compared with the traditional linear prediction model, the prediction precision is increased further, and the result is more ideal [7-10]. Since online public opinion is characterized by people's participation, users have their own preferences and ideas, making online public opinion have time-varying and chaos. The current machine learning algorithm ignores the chaotic characteristics of the online public opinion, thus the established model cannot comprehensively and accurately describe the online public opinion change trend, and the prediction precision remains to be further improved [11].

In view of the chaos of online public opinion change, phase space reconstruction (PSR) and SVR are combined together, forming a online public opinion trend prediction model based on PSR - SVR, and through the simulation experiment, the validity of the PSR–SVR is tested.

2. Phase Space Reconstruction and Support Vector Regression

2.1. Phase Space Reconstruction

Phase space reconstruction theory is the basis of chaotic time sequence prediction, and the main idea is: any component evolution of the system and the interaction is determined by other components, and its related components' information is hidden in the evolution process, so through the analysis of a certain component's time sequence, the dynamic characteristics of the original system can be understood, with the extraction and restore of the original system [12].

For time sequence of online public opinion, x (t), t = 1, 2,... , N, by selecting an appropriate embedding dimension m and time delay τ, time sequence can be reconstructed, get a set of multi-dimensional vector sequence of the formula (1), so as to excavate the time sequence hidden in the online public opinion, and restore online public opinion motive power system.

$$X(t) = \{x(t), x(t+\tau), \mathrm{n}, x[t+(m-1)\tau]\} \tag{1}$$

Wherein, $M=N-(m-1)\tau$.

This paper uses the method of mutual information and G - P method to calculate and determine the online public opinion τ and m.

2.2. Support Vector Regression

Support vector machine is a kind of machine learning algorithm based on statistical learning theory, to find the best compromise between the model complexity and the ability to learn, in order to get the best generalization ability [13]. The SVR regression estimate function is

$$f(x) = w \cdot \varphi(x) + b \tag{2}$$

Wherein, w is weight vector, b is the bias vector.

Making the predictive expected risk function minimum

$$\min J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^* + \xi_i) \tag{3}$$

Constraint condition is:

$$\begin{cases} y_i - w \cdot \varphi(x) - b \le \varepsilon + \xi_i \\ w \cdot \varphi(x) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, \mathrm{n}, n \end{cases} \tag{4}$$

Wherein, ξ_i, ξ_i^* is relaxing factor, C is penalty factor.

Throughout introducing Lagrange multiplier, the above optimization problem is changes as typical convex quadratic optimization problem, namely,

$$\begin{aligned} L(w,b,\xi,\xi^*,\alpha,\alpha^*,\gamma,\gamma^*) &= \frac{1}{2}\|w\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &-\sum_{i=1}^{n}\alpha_i(\xi_i + \varepsilon - y_i + f(x_i)) - \sum_{i=1}^{n}\alpha_i^*(\xi_i^* + \varepsilon - y_i + f(x_i)) \\ &-\sum_{i=1}^{n}(\xi_i\gamma_i - \xi_i^*\gamma_i^*) \end{aligned} \tag{5}$$

Wherein, α_i and α_i^* represent Lagrange multiplier.

$$\begin{aligned} W(\alpha,\alpha^*) &= -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\varphi(x_i), \\ &\varphi(x_j)) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varepsilon \end{aligned} \tag{6}$$

Constraint condition is:

$$\begin{cases} w = \sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)x_i \\ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{7}$$

Upon solving the practical questions, it only needs to make use of support vector machine to get the solution, so the regression estimation function is

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)(\varphi(x_i), \varphi(x)) + b \tag{8}$$

Kernel function $k(x_i, x)$ is adapted to replace $(\varphi(x_i), \varphi(x))$, which can avoid curse of dimensionality, so

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{9}$$

This paper selects radial basis function kernel as kernel function of SVR, finally regression function of SVR:

$$f(x) = \sum_{i=1}^{N}\alpha_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) + b \tag{10}$$

Wherein, σ is the width of radial basis function kernel [13].

3. PSR-SVR Online Public Opinion Trend Prediction Model

(1) First of all, based on heuristic search spider, the network data are fetched with impurity elimination and data storage.

(2) Subtopic-field segmentation is conducted. By using the method of automatic threshold calculation, hierarchical topic tree is established, with the hierarchical clustering analysis, and the purity and F value is used for clustering quality evaluation.

(3) Lastly, media attention as well as the public interest is calculated. Based on the proportion of a given balance factor, comprehensive attention is calculated, and then hot judgment and scraping are conducted. Then, data aggregation software is used for data aggregation, to generate discrete time sequence.

(4)The SVR parameters are set.

(5) The delay time τ is obtained by using the mutual information method, and G - P method is used to calculate embedding dimension m.

(6) τ and m are selected to conduct phase space reconstruction of chaotic time sequence, and is divided into training set and test set.

(7) The training set is input in the SVR to learn, and particle swarm optimization is used to select the SVR parameters to build an online public opinion prediction model.

(8) Online public opinion prediction model is established to predict the test set, and analyze its prediction performance.

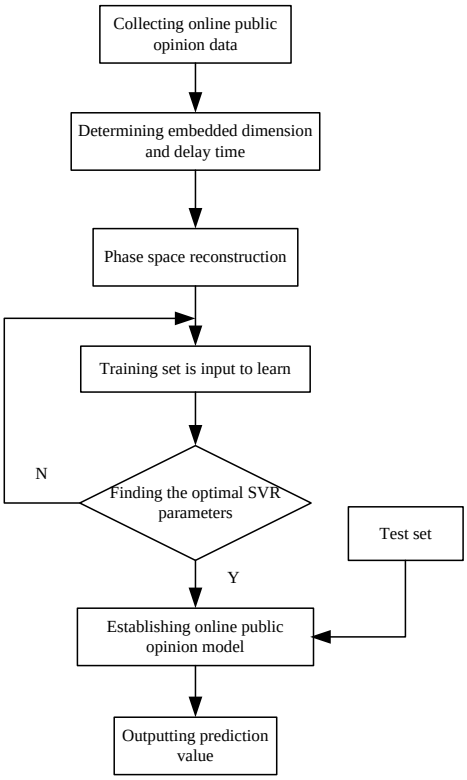

Figure 1. Online Public Opinion Prediction Procedure of PSR-SVR

4. Simulation Experiment

4.1. Data Source

In P4 3.0 G CPU, 2 G RAM, the operating system is Windows 2000 environment, and the algorithm is realized through the VC + + programming. "Changchun baby onboard theft" is chosen as the source event of the online public opinion. Since Tianya Community is top 1 in the rank of “top 100 global Chinese BBS” jointly issued by PhoenixNet and iResearch Consulting Group, with the advantages of popularity and influence degree, and its data are representative, so data from Tianya BBS are chosen as data source of online public opinion. From March 4, 2013, 10 am in Tianya, when the first source of "Changchun baby onboard theft" appeared, to March 8, 2013, the number of posts on "Changchun baby onboard theft "in the total about 96 hours are collected, as shown in Figure 2. The data are divided into two parts, the first 66 data as the training set, the remained 30 data as the test set.

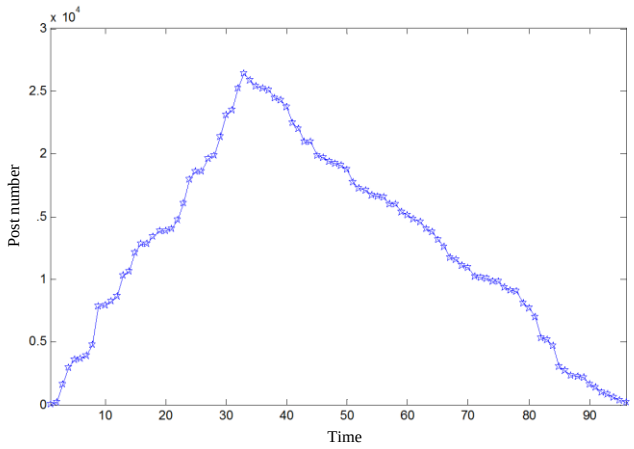

Figure 2. Collecting Online Public Opinion Data

4.2. Contrast Model and Evaluation Standard

ARIMA, SVR (without phase space reconstruction), PSR - BPNN are used as contrast models. The root mean squared error (RMSE) and the mean percentage average error (MPAE) are used as evaluation standard for model. They are defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - x_t)^2} \tag{11}$$

$$MAPE = \frac{\sum_{t=1}^{n}\left|(x_t - x_t)/x_t\right|}{n} \times 100 \tag{12}$$

Wherein, x_t and x_t are actual and model prediction values, respectively, n as the sample size.

4.3. Pretreatment of Online Public Opinion

From Figure 2, it can be concluded that the range of online public opinion change is larger, in order to avoid the data with a large range covers the data with a small range. And the value of SVR kernel function depends on the inner product of the characteristic vector, too large value will adversely affect the process of training, therefore, before the data is input to the SVR, the normalized processing shall be conducted. The normalization formula is:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{13}$$

Wherein, x' is the normalized value, $x_{\max}$ and $x_{\min}$ are respectively the maximum and the minimum value.

4.4. Phase Space Reconstruction of Online Public Opinion Data

4.4.1. Delay Time Calculated by Mutual Information Method

(1) Building a two-dimensional phase diagram of time sequence of online public opinion {x (t)}, and making (x,y)=[x(t),x(t+τ), τ=1.

(2) In two-dimensional phase diagram, drawing the rectangular box of the attractor, and making the rectangular box divided into sub-boxes, *x0* and *y0* are the starting point

of the sub-box, and Δx and Δy are the length of the sub-boxes x and y directions, *Mx* and *My* are the number of sub-boxes on x and y directions.

(3) If $x0 \le x(i) \le x0+\Delta x$, $y0 \le y(j) \le j0+\Delta y$, i,j=1,2,…,N, then the point [x(i),y(j)] is in rectangular box, and satisfys $(k-1)\Delta x \le x(i)-x0 \le k\Delta x$, $(l-1)\Delta y \le y(i) \le y(i)-y0 < k\Delta y$, k=1,2,…,M, l=1,2,…,M, then the point [x(i),y(j)] is in the sub-box Δk,j, and record for once, searching all the points, the number of falling into a (k, l) sub-box is Nxy, and the point of being in k - 1 to k sub-boxes is Nx, the number of being in the l - 1 to 1 boxes is Ny. p[x(i)]=Nx/N, p[y(i)]=Ny/N, p[x(i),y(i)]=Nxy/N, N is all sampling points, set into the different formulas (14) ~ (16), to get the mutual information function value of the delay time as τ, I (x, y).

$$H(x) = -\sum_{i=1}^{q} P(x_i) \log P(x_i) \tag{14}$$

$$H(x,y) = -\sum_{i,j} P(x_i, y_i) \log P(x_i, y_i) \tag{15}$$

$$I(x,y) = H(X) + H(Y) - H(X,Y) \tag{16}$$

Wherein, H (X) represents the uncertain degree of X, $P(xi)$ is the incidence of *xi*, *q* is the total number of state, $H(X,Y)$ is the joint entropy of X and Y, $P(xi,yi)$ is joint probability of the event of *xi* and *yi*.

(4) Making $\tau=\tau+1$, returning to step (2).

The mutual information function change curve of online public opinion time sequence is as shown in Figure 3. From Figure 3, it can be seen, when $\tau=3$, mutual information function achieves the minimum, so the online public opinion time sequence $\tau=3$.

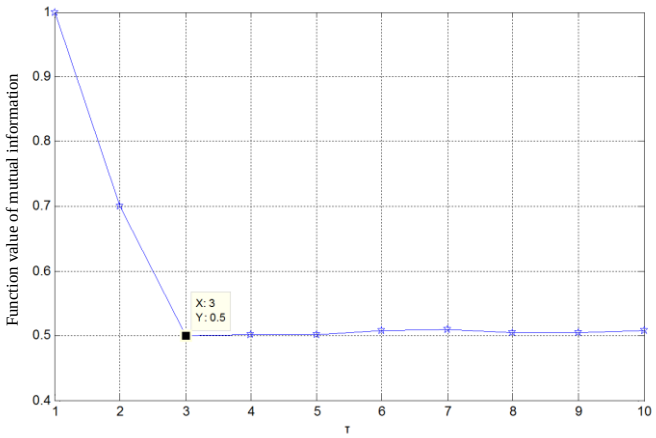

Figure 3. Calculation of Delay Time of Online Public Opinion

4.4.2. Embedded Dimension Chosen by G-P Method

(1) According to the mutual information, τ=3 is obtained, the initial value of embedded dimension is *m*=1.

(2) Selecting proper critical distance *r;* according to the formula (17), calculating $C_n(r)$. Vector distance is calculated by ∞ norm, namely the maximum differential component of two vectors are vector distance.

$$C_n(r) = \frac{1}{M^2} \sum_{i,j=1}^{M} \theta\left[r - \left\|X(i) - X(j)\right\|\right] \tag{17}$$

Wherein, *M* is the number of phase point, *r* is critical distance, θ is Heaviside unit function.

Least square method is used to fit tangential path in $\log C(r) n \sim \log r$ curve, and the slope of straight line is correlation dimension D.

Correlation dimension of online public opinion time sequence under different embedded dimensions is as shown in Figure 4. From the Figure 4, it shows when the embedding dimension m = 5, correlation dimension reaches saturation correlation dimension, and this shows that the optimal online public opinion time sequence m = 5.

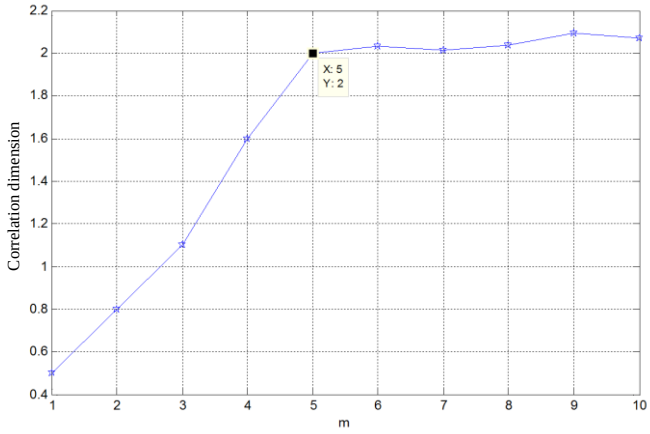

Figure 4. Calculation of Embedded Dimension of Online Public Opinion

4.4.3. Chaotic Judgment of Online Public Opinion

By computing, the average period of online public opinion time sequence is obtained p=1, embedded dimension m=5, delay time τ=3. Least square method is used to fit straight line, the slope is the maximum Lyapunov index, to getλmax=0.00152>0, which shows that the online public opinion time sequence has weakly chaotic characteristics.

4.4.4. One-Step Prediction

Because the optimal embedded dimension is 5, the original training sample is 66-5 = 61. Firstly, the training containing 61 data is conducted for one step prediction, and then for each time, a true value is added for one-step prediction, to calculate a total of 30 prediction values, and calculate the corresponding RMSE and MAPE. The structure of the BP neural network is 5-11-1; throughout the particle swarm algorithm, the optimal SVR is obtained, C = 100, σ= 1.715, ARIMA model selects ARIMA (3,2,2). The prediction results of all models on the test set of online public opinion are as shown in Figure 5, and their corresponding RMSE and MAPE are shown in Table 1.

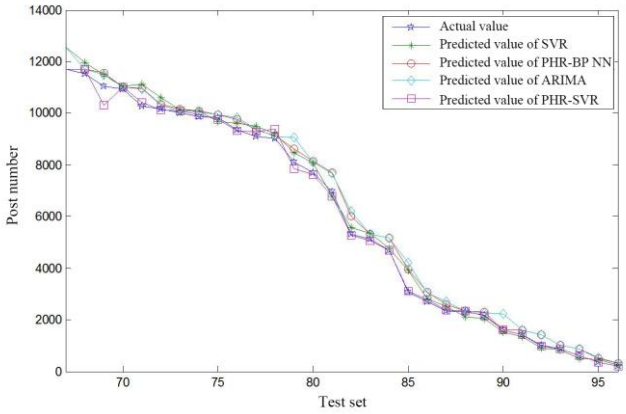

Figure 5. One-Step Prediction Result of Online Public Opinion

Table 1. Contrast of One-Step Prediction Error of All Models

Model	*RMSE*	*MAPE*
ARIMA	468.274	14.05%
PSR-BPNN	400.997	11.71%
SVR	350.187	6.58%
PSR-SVR	184.501	1.96%

Analysis is conducted based the results in Table 1 and Figure 5, and it can be concluded:

(1) Compared with ARIMA, PSR-SVR online public opinion prediction precision increases significantly, which is mainly due to the ARIMA that is unable to capture the nonlinear change characteristics of online public opinion time sequence, and PSR - SVR uses SVR's nonlinear prediction ability to effectively improve the prediction precision of the online public opinion.

(2) Compared with SVR, the prediction error value of PSR - SVR online public opinion is smaller, the prediction value and the real value are very close, which is mainly because the PSR - SVR explores in the information on time sequence of online public opinion by means of PSR, and it can be more accurately and comprehensively describe online public opinion tendency, get more reliable prediction results, and further improve the precision of online public opinion changes.

(3) Compared with PSR - BPNN, PSR - SVR prediction results are relatively Table, and the RMSE and MAPE value of prediction results is far less than the PSR - BPNN, which is mainly because the SVR well overcomes the BP neural network fitting, local minimum and problems that network parameters is difficult to determine, with stronger generalization ability and higher prediction precision.

4.4.5. Multi-Step Prediction

Online public opinion prediction time generally requires larger lead. Using the one-step prediction (that is, predicting the online public opinion next hour), cannot effectively reflect the trend of online public opinion, nor make effective and timely response to some negative online public opinions, therefore, it is necessary to expand one-step prediction to multi-step prediction method, then the multi-step prediction method is used to predict online public opinion in the future 24 hours. The contrast between real value and

prediction value of prediction results of all models is shown in Figure 6. The RMSE and MAPE are shown in Table 2.

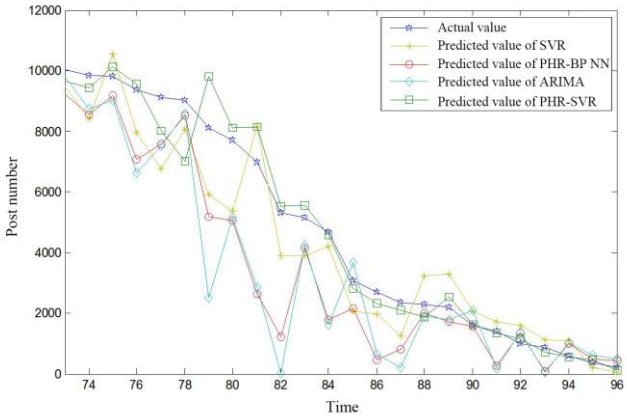

Figure 6. Contrast of Prediction Results of Multi-Step Prediction

Table 2. Contrast of Multi-Step Prediction Performance

Model	*RMSE*	*MAPE*
ARIMA	2216.429	39.61%
PSR-BPNN	1848.875	32.19%
SVR	1169.536	21.62%
PSR-SVR	680.8821	9.00%

From the Figure 6 and Table 2, it can be seen the multi-step prediction accuracy of ARIMA, SVR, PSR - BPNN online public opinion is low, the error is quite high, the prediction results are not reliable, the practical application value of the prediction results is low. PSR - SVR prediction error is less than the contrast model obviously, and PSR – SVR's prediction of online public opinion change trend is more accurate, the prediction performance is superior to the contrast model, and the prediction results have great practical value.

5. Conclusion

The online public opinion is affected by various factors, and is characterized by time-varying, chaos. It is a kind of complex changes system, and the traditional prediction algorithm is difficult to establish accurate prediction model. According to the chaotic characteristics of online public opinion changes, by using chaos theory and SVR, a model of online public opinion prediction based on PSR – SVR was built. Results show that compared with contrast model, PSR - SVR improves the prediction precision of online public opinion, prediction results are more sTable, and it more accurately describes the complex change trend of the online public opinion. The prediction results are helpful to correctly understand the development of online public opinion, thus helping to scientifically guide and manage various online public opinion transmission platforms, to promote the work of building a harmonious society.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (51309098).

Reference

[1] J. Hu, Z. Gao and W. Pan, "Multi-angle Social Network Recommendation Algorithms and Similarity Network Evaluation", Journal of Applied Mathematics, vol. 2013, **(2013)**.
[2] Z. Lv, T. Yin, Y. Han, Y. Chen and G. Chen, "WebVR—web virtual reality engine based on P2P network", Journal of Networks, vol. 6, no. 7, **(2011)**, pp. 990-998.
[3] J. Yang, B. Chen, J. Zhou and Z. Lv, "A portable biomedical device for respiratory monitoring with a stable power source", Sensors, **(2015)**.
[4] Z. Dongfang, "Fusion FS: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems", Big Data (Big Data), 2014 IEEE International Conference on. IEEE, **(2014)**.
[5] S. Dang, J. Ju, D. Matthews, X. Fen and C. Zuo, "Efficient solar power heating system based on lenticular condensation", Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on, 26-28 April **(2014)**.
[6] X. Zhang, Y. Han, D. Hao and Z. Lv, "ARPPS：Augmented Reality Pipeline Prospect System", 22th International Conference on Neural Information Processing (ICONIP), Istanbul, Turkey. In press, **(2015)**.
[7] J. Hu and Z. Gao, "Distinction immune genes of hepatitis-induced heptatocellular carcinoma", Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.
[8] W. Ke, "Overcoming Hadoop Scaling Limitations through Distributed Task Execution".
[9] Z. Su, X. Zhang and X. Ou, "After we knew it: empirical study and modeling of cost-effectiveness of exploiting prevalent known vulnerabilities across IAAS cloud", Proceedings of the 9th ACM symposium on Information, computer and communications security. ACM, **(2014)**.
[10] G. Bao, L. Mi, Y. Geng and K. Pahlavan, "A computer vision based speed estimation technique for localizing the wireless capsule endoscope inside small intestine", 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)**.
[11] W. Gu, Z. Lv and M. Hao, "Change detection method for remote sensing images based on an improved Markov random field", Multimedia Tools and Applications, **(2016)**.
[12] Y. Wang, Y. Su and G. Agrawal, "A Novel Approach for Approximate Aggregations Over Arrays", In Proceedings of the 27th international conference on scientific and statistical database management, ACM, **(2015)**.
[13] J. Hu and Z. Gao, "Modules identification in gene positive networks of hepatocellular carcinoma using Pearson agglomerative method and Pearson cohesion coupling modularity", Journal of Applied Mathematics, vol. 2012, **(2012)**.
[14] X. Li, Z. Lv, J. Hu, L. Yin, B. Zhang and S. Feng, "Virtual Reality GIS Based Traffic Analysis and Visualization System", Advances in Engineering Software, **(2015)**.
[15] Z. Lv, C. Esteve, J. Chirivella and P. Gagliardo, "Clinical Feedback and Technology Selection of Game Based Dysphonic Rehabilitation Tool", 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2015), IEEE, **(2015)**.
[16] W. Ke, "Next generation job management systems for extreme-scale ensemble computing", Proceedings of the 23rd international symposium on High-performance parallel and distributed computing. ACM, **(2014)**.
[17] L. Tonglin, "Distributed Key-Value Store on HPC and Cloud Systems", 2nd Greater Chicago Area System Research Workshop (GCASR), **(2013)**.
[18] T. Su, W. Wang, Z. Lv, W. Wu and X. Li, "Rapid Delaunay Triangulation for Random Distributed Point Cloud Data Using Adaptive Hilbert Curve", Computers & Graphics, **(2015)**.
[19] Z. Lu, C. Esteve, J. Chirivella and P. Gagliardo, "A Game Based Assistive Tool for Rehabilitation of Dysphonic Patients", 3rd International Workshop on Virtual and Augmented Assistive Technology (VAAT) at IEEE Virtual Reality 2015 (VR2015), Arles, France, IEEE, **(2015)**.
[20] Z. Chen, W. Huang and Z. Lv, "Uncorrelated Discriminant Sparse Preserving Projection Based Face Recognition Method", Multimedia Tools and Applications, **(2016)**.

Author

Song Liangong, Born in 1975, MSc, a lecturer at North China University of Water Resources and Electric Power in china, Beng from China Agricultural University in 1999, MSc in Information Management Study form Beijing Forestry University in 2005. Research interest is in the area of Theory and Application of Computer, Principle and Application of Database, Information Management.